A FUNCTIONAL INVESTIGATION OF THE FREQUENTLY MUTATED SPLICING FACTOR U2AF1 IN NON-SMALL CELL LUNG CANCER

BY
NORIKO ISHISOKO

UNDER THE SUPERVISION OF
DR. MAXIMILIAN DIEHN
DEPARTMENT OF RADIATION ONCOLOGY

STANFORD UNIVERSITY

MAY 2014
A FUNCTIONAL INVESTIGATION OF THE FREQUENTLY MUTATED SPICING FACTOR U2AF1 IN NON-SMALL CELL LUNG CANCER

NORIKO ISHISOKO

STANFORD UNIVERSITY

12 MAY 2014

APPROVED:

MAXIMILIAN DIEHN, M.D., PH.D.
RESEARCH ADVISOR
DEPARTMENT OF RADIATION ONCOLOGY

APPROVED:

RUSS ALTMAN, M.D., PH.D.
FACULTY READER
DEPARTMENT OF BIOENGINEERING

APPROVED:

KARL DEISSEROTH, M.D., PH.D.
CHAIR FOR UNDERGRADUATE EDUCATION
DEPARTMENT OF BIOENGINEERING

THIS HONORS THESIS HAS BEEN SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE IN BIOENGINEERING WITH DISTINCTION AT STANFORD UNIVERSITY.
Preface

This project would not have been possible without the guidance of Dr. Maximilian Diehn, Dr. Scott Bratman, Luke Lee, and Dr. Ash Alizadeh, all from the Stanford Cancer Institute. Thank you to these practicing doctors for taking time out of their important lives to help an undergrad. Special thanks to Eric Kildebeck and Dr. Matthew Porteus for their help with the TALENs part of this project, Dr. Russ Altman for being my Bioengineering advisor and reader of my thesis, and to Teri Hankes for her support throughout my undergraduate career in the Department of Bioengineering. The most special thank you has been saved for my parents, Nelia and Tetsuo Ishisoko, without whom I would be nothing.
# Table of Contents

Abstract .................................................................................................................. iv

I. Introduction ............................................................................................................. 1

II. Background .......................................................................................................... 2
   - Alternative Splicing and Mutant U2AF1 ......................................................... 2
   - ROS1 Rearrangements and Kinase Inhibitors ................................................. 5
   - HCC78 ............................................................................................................. 6
   - Genomic Editing and TALEN Technology ...................................................... 8
   - Next Generation Sequencing and RNA-seq ................................................... 10

III. Methods ............................................................................................................. 14
   - siRNA Knockdown Assay ............................................................................ 14
   - siRNA Knockdown Assay RNA Transcription Analysis ............................... 14
   - TALEN Design and Synthesis ..................................................................... 14
   - Donor DNA Design and Synthesis .............................................................. 15
   - Transfection ................................................................................................. 17
   - Drug Assays .................................................................................................. 17
   - RNA-seq ....................................................................................................... 18

IV. Results .............................................................................................................. 20
   - siRNA Knockdown of U2AF1 and Resulting RNA Expression ..................... 20
   - TALENs and Donor DNA Transfection ....................................................... 21
   - Drug Assays .................................................................................................. 23
   - RNA-seq ....................................................................................................... 23

V. Discussion .......................................................................................................... 23
   - siRNA Knockdown Assay ............................................................................ 23
   - TALENs and Donor DNA Transfection ....................................................... 24
   - Drug Assays and RNA-seq ......................................................................... 24

VI. Conclusions and Future Work ......................................................................... 25

VII. References ...................................................................................................... 26
List of Figures

Figure 1: Examples of Alternative Splicing................................................. 3
Figure 2: The Spliceosome ......................................................................... 4
Figure 3: Mutated U2AF1 Gene.................................................................. 5
Figure 4: Breakdown of NSCLC................................................................. 6
Figure 5: SLC34A2, ROS1, and SLC34A2-ROS1 .................................... 7
Figure 6: DSB Repair .................................................................................. 8
Figure 7: TALENs vs. ZFNs..................................................................... 10
Figure 8: High-Throughput vs. Sanger Sequencing .................................. 12
Figure 9: A Modified Tuxedo Protocol ..................................................... 13
Figure 10: Designed TALEN Pairs ............................................................ 15
Figure 11: Designed Donor DNA ............................................................... 16
Figure 12: cDNA Insertion into the Genome ............................................. 16
Figure 13: Developing a stable cell line ..................................................... 17
Figure 14: siRNA Knockdown Expression Data ....................................... 21
Figure 15: Flow Cytometry Data of Transfected Cells .............................. 22
Abstract

RNA splicing is a critical step in manufacturing most human proteins. Regulating the splicing machinery is crucial for normal development, and aberrant splicing can result in diseases such as cancer. Recent studies have uncovered a recurrent mutation of the splicing factor U2AF1 in several human cancers including lung cancer. The lung cancer cell line HCC78 is the only cancer cell line known to harbor this mutation, and it also has a gene fusion involving the ROS1 gene that genetically separates it from other types of lung cancer. This study sets out to both examine splicing effects of the mutated U2AF1 in lung cancer and to clarify the relationship between mutant U2AF1 and the ROS1-fusion, two rare genetic alterations that have been observed to occur together in lung cancers at a significantly higher than expected frequency. Specifically, by genomically editing the HCC78 cell line and repairing the U2AF1 point mutation, an appropriate point of comparison has been created using transcription activator-like effector nucleases (TALENs). Use of TALENs allows for targeting sequence-specific locations in the genome for double-stranded breaks. Taking advantage of endogenous DNA repair machinery, a designed sequence has been inserted into the genome of these HCC78 cells to repair the U2AF1 point mutation.
I. Introduction

In the central dogma of molecular biology, eukaryotic precursor messenger RNAs (pre-mRNAs) are processed via splicing in the nucleus before they are exported to the cytoplasm for translation into proteins. Different pre-mRNA splicing events can lead to different proteins, even though they all have been transcribed from the same gene. This process, called alternative splicing, is carried out by the spliceosome and is a naturally occurring, regulated part of development and daily activity. The U2AF1 gene codes for an essential component of the spliceosome. However, current literature does not agree on the consequences of a mutation in U2AF1: Yoshida et al. concluded that mutant U2AF1 leads to a loss of gene function and to less splicing events.\(^1\) A few months later, Graubert et al. indicated that mutating U2AF1 leads to gain-of-function activity and thus to more splicing events.\(^2\) Both studies involved the use of exogenous expression, \textit{i.e.} use of transient plasmids for mutant expression, to come to their conclusions. These opposing conclusions indicate that further study is needed to clarify the effects of the U2AF1 mutation.

Lung cancer is the leading cause of cancer deaths in the United States,\(^3\) and it is comprised of a diverse set of tumors. Lung cancers are divided into subtypes based on histological appearance and protein expression and more recently have been further sub-categorized by specific somatic, genetic alterations. Non-small cell lung cancer (NSCLC) is the cancer of interest in this study. Within the last few years, recurrent rearrangements involving ROS1 have been identified as a molecular subtype of the histological subtype lung adenocarcinoma.\(^4\) ROS1 is a receptor tyrosine kinase and is involved in cell signaling. The lung adenocarcinoma cell line HCC78 harbors a ROS1-rearrangement and has demonstrated sensitivity to the kinase inhibitors crizotinib and TAE684.\(^4\) A statistically significant co-occurrence of ROS1-rearrangements with the U2AF1
mutation has been found by sequencing analysis of hundreds of lung tumor samples.\textsuperscript{5} HCC78 also has the heterozygous U2AF1 mutation. Because of their statistically significant co-occurrence, it has been hypothesized that there is an unrecognized interaction between these two rare, genetic alterations.

The purpose of this study is therefore to investigate the effects of the mutant U2AF1 gene on gene expression and global splicing events, as well as this mutation’s effect on cell proliferation. Also being examined is whether or not the U2AF1 mutation modulates sensitivity of ROS1-rearranged cells’ sensitivity to kinase inhibitors. The lung adenocarcinoma cell line HCC78 contains a ROS1-rearrangement as well as the U2AF1 point mutation of interest. This cell line will be genomically edited to contain only wild type alleles of U2AF1, and the resulting stable cell line will serve as the point of comparison (cf. a transient, exogenously expressed mutation) for this study.

II. Background

Alternative Splicing and Mutant U2AF1

U2AF1, also known as U2AF35, is a component of the spliceosome. The spliceosome is a complex of RNA and proteins that in its simplest function remove introns from pre-mRNA.\textsuperscript{6} Given that about 24,000 protein-coding genes are translated into 100,000 different proteins, splicing explains how one gene can produce multiple proteins, called isoforms.\textsuperscript{7} Besides the canonical removal of introns and ligation of exons, alternative splicing includes processes such as those depicted in Figure 1: exon skipping, alternative 3’ or 5’ splice site (SS) selection, and intron retention. Exon skipping removes an exon flanked by intronic sequences. Alternative 3’ and 5’ SS selection occurs when two or more SSs are
recognized at the end of an exon, resulting in partial splicing of said exon. Intron retention happens if an intron remains in the mature mRNA transcript.\(^7\)

![Figure 1](image.png)

**Figure 1.** Shown here are some examples of alternative splicing (AS). Dashed lines encompass segments of DNA that will be spliced out. Constitutive exons are shown in light blue, and alternatively spliced regions are shown in purple. Introns are represented by solid lines. Canonical splicing is indicated by the dashed lines below the gene. AS events are indicated by the dashed lines above the gene. (a) depicts exon skipping, which accounts for nearly 40% of AS events in higher eukaryotes. (b) and (c) depict alternative 3′ and 5′ splice site (SS) selection, respectively. Alternative 3′ SS selection accounts for 18% and alternative 5′ SS selection accounts for 8% of all AS events in higher eukaryotes. (d) depicts intron retention, the rarest AS event (<5% of all AS events). Figure adapted from Keren’s “Alternative splicing and evolution: diversification, exon definition, and function.”

Alternative splicing events are not rare and are an important mechanism in enhancing transcriptome and proteome diversity in eukaryotes. Aberrant splicing has been indicated in multiple human diseases including cancer.\(^8\)

As depicted in **Figure 2**, the spliceosome is composed of many proteins and five small nuclear ribonucleoproteins (snRNPs) named U1, U2, U4, U5, and U6 that together bind to and splice pre-mRNA. The snRNPs are composed of small
nuclear RNAs that assemble with proteins. U2AF1 (U2 auxiliary factor 1) is part of a heterodimer that binds the 3’ splice site and helps recruit the U2 snRNP to the branch site.\(^7\)

**Figure 2.** (a) illustrates the conserved RNA-recognition signals: exon-intron junctions (5’ SS and 3’ SS), the branch site sequence, and the polypyrimidine tract (PPT). U2AF1 protein binds the 3’ SS and helps recruit the U2 snRNP to the branch site. (b) shows the main components of the spliceosome, the snRNPs. Unlabeled orange ovals are other components of the spliceosome, of which U2AF1 is one. Blue boxes are exons, and solid lines are introns. Figure adapted from Keren’s “Alternative splicing and evolution: diversification, exon definition, and function.”

U2AF1 protein mutations are most often substitution missense changes, the two most common being S34F and Q157P (see Figure 3).\(^9\) These mutations fall in zinc finger domains, suggesting a possible functional change. The U2AF1 S34F mutation will be the focus of this study, and its function has yet to be determined. Current correlative data indicate that various splicing factors are upregulated or downregulated in cancers.\(^8\)
ROS1 Rearrangements and Kinase Inhibitors

With the advent of the continually decreasing costs of genomic sequencing, tumors have been sequenced and classified genotypically. Lung adenocarcinoma is the most common form of NSCLC and has been further divided into different categories of typically mutually exclusive mutations, as shown in Figure 4. With different types of tumors, the hope is to discover treatments that work for specific tumors. For instance, tumors with epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma kinase (ALK) rearrangements have shown a clinical response to certain kinase inhibitors. Specifically, gefitinib and erlotinib are part of targeted therapies for patients with EGFR mutations, and similarly, crizotinib is for ALK-rearrangements. ROS1 is a gene that codes for a receptor tyrosine kinase, as is ALK and EGFR. Recently, ROS1 rearrangements have been identified as another subset, and crizotinib has also been shown to exhibit a clinical response, paving the way for another targeted therapy for a different set of patients. Although the percentage of patients with these specific
mutations is small, the number of potentially affected patients is in the thousands, since lung cancer afflicts so many individuals. One current impediment to ubiquitous personalized genomic-specific treatments (besides existence) is screening, since the current genotyping gold standard of fluorescence in situ hybridization (FISH) is impractical for rapid screening and is not cost effective.\textsuperscript{10}

HCC78

HCC78 was chosen for this study because it contains a ROS1 rearrangement and U2AF1 mutation. It is a cell line developed from the pleural effusion of a 65-year-old man with NSCLC adenocarcinoma of the lung.\textsuperscript{11} This U2AF1 mutation

\textbf{Adenocarcinoma}

\textit{Figure 4.} Shown (top) are the histologic and molecular subtypes of non-small-cell lung cancer (NSCLC). Adenocarcinoma, the most common subtype of NSCLC, has further been divided into somatic genomic alterations (bottom). ROS1 rearrangements are one specific genomic alteration. Figure adapted from Jänne’s “ROS1 Rearrangements in Lung Cancer: A New Genomic Subset of Lung Adenocarcinoma.”
affects one nucleotide on one allele of the U2AF1 gene. The wildtype allele contains a cytosine (C) where the mutant contains a thymine (T) in the 101\textsuperscript{st} nucleotide of the coding sequence for this gene (the first nucleotide is the adenine (A) from the start codon (ATG) in this gene).

In HCC78, ROS1 is fused to a nearby gene named solute carrier family 34 (sodium phosphate), member 2, or SLC34A2 for short.\textsuperscript{12} This SLC34A2-ROS1 rearrangement translates to two main isoforms—long and short\textsuperscript{13} (see Figure 5). The long form fuses SLC34A2 exon 4 to ROS1 exon 32, and the short form fuses SLC34A2 exon 4 to ROS1 exon 34.\textsuperscript{14} Currently, the functional relevance of these two isoforms is unclear.

**Figure 5.** SLC34A2, ROS1, and SLC34A2-ROS1 genes are depicted. There are two main isoforms of SLC34A2-ROS1: long and short. The red area on the long isoform indicates a retained transmembrane domain. Figure adapted from Rikova’s “Global Survey of Phosphotyrosine Signaling Identifies Oncogenic Kinases in Lung Cancer” and Stumpova’s “Zeroing in on ROS1 Rearrangements in Non-Small Cell Lung Cancer.”
Genomic Editing and TALEN Technology

The most critical part of this study is the engineering of a stable HCC78 cell line with a repaired point mutation. Methods such as homologous gene targeting, transposases, site-specific recombinases, meganucleases, and integrating viral vectors have been used to insert and knockout genes. Limitations of these methods include targeting DNA recognition sequences that cannot be modulated and nonspecific integration. With zinc-finger nucleases (ZFNs), however, specific DNA sequences can be targeted for double-stranded breakage, and then either nonhomologous end joining (NHEJ) or homologous recombination (HR) repairs the break, as shown in Figure 6. NHEJ joins the two broken ends and may result in loss of nucleotides, and HR uses the homologous chromosome as a template for repair. Both are naturally occurring processes for DNA repair.

![Figure 6](image)

**Figure 6.** When a nuclease induces a double-stranded break (DSB), it can be repaired via non-homologous end-joining (NHEJ) or homology directed repair (HDR). With NHEJ, mutations typically arise (left). Normally in the cell, homologous recombination (HR) repairs the DNA to the sequence that it should be (center). With a donor template that has enough homology, HR can lead to an insertion (right). This is the concept that drives the creation of a stable cell line in this study. Figure adapted from Joung’s “TALENs: a widely applicable technology for targeted genome editing.”

ZFNs are comprised of a cleavage part—a FokI nuclease that cuts nonspecifically—and a DNA recognition part—Cys$_2$His$_2$ zinc finger DNA
binding domains that recognize specific nucleotide triplets. However, triplet recognition limits the number of sequences that can be targeted, compared to that of transcription activator-like effector nucleases (TALENs), which recognize single nucleotides (see Figure 7).\textsuperscript{16} TALENs are similarly comprised of a FokI nuclease and a customizable DNA-binding domain.\textsuperscript{17} Here the binding domain is comprised of a TALE polypeptide that consists of 34-amino acid modules. These modules are variable on the 12\textsuperscript{th} and 13\textsuperscript{th} residues (called the repeat variable diresidue, or RVD), which enables recognition of a specific nucleotide based on a simple code.\textsuperscript{18} Putting these modules together allows for recognition of specific sequences. Both ZFNs and TALENs require dimerization to initiate site-specific cleavage. In this study, TALENs will be utilized to repair the HCC78 U2AF1 point mutation.
As sequencing technology progresses, so do the methods that result from it. High-throughput sequencing has led to the RNA-seq method: using massively parallel cDNA sequencing and bioinformatics to analyze the results.
transcriptome analysis can reveal changes in gene expression and evidence for alternative splicing.\textsuperscript{19} In RNA-seq, total RNA is transcribed into cDNA with adaptors that enable millions of fragments to be sequenced at a time. This is different from the more typical Sanger sequencing, where only one strand can be sequenced linearly at a time (see Figure 8). Gigabytes of data are produced, and in this study a modified Tuxedo Protocol\textsuperscript{20} (Figure 9) will be used to analyze it. Running Spliced Transcripts Alignment to a Reference (STAR) software\textsuperscript{21} will first create a reference genome using available genomic data from Illumina. Then Cufflinks will align the sequencing reads to this reference genome. Cuffmerge will create an experiment specific transcriptome, and with this, Cuffdiff will analyze the samples for differential gene expression.
Figure 8. (a) Workflow for Sanger sequencing. Genomic DNA is fragmented, then a small portion is cloned into a vector that is transformed into E. coli. E. coli make many copies of this plasmid DNA and it is isolated. The sequencing reaction takes advantage of ddNTP-terminated, dye-labeled products that can be separated by size via electrophoresis, and the resulting sequence is recorded using a four-channel emission spectrum. (b) Workflow for high-throughput sequencing. Genomic DNA is fragmented and adaptor sequences are ligated to the fragments. These fragments are then spatially immobilized to form PCR colonies or “polonies.” In a sequential fashion, each nucleotide of each fragment’s sequence is determined using cycles of enzymatic extension and imaging-based data collection. Figure adapted from Shendure & Ji’s “Next-generation DNA sequencing.”
Figure 9. A modified Tuxedo Protocol. Instead of using TopHat as the original protocol suggests, STAR has been used to align the reads. Condition A and Condition B represent different conditions, for instance experimental vs. control. Figure adapted from Trapnell et al.’s “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.”
III. Methods

siRNA Knockdown Assay

To investigate if there is any effect of knocking down all alleles of the U2AF1 gene, a small interfering RNA (siRNA) knockdown assay was performed on HCC78 cells. Using commercial siRNA designed to target the 3’UTR of U2AF1, the U2AF1 gene was knocked down (ON-TARGETplus U2AF1 siRNA, ThermoScientific [#J-012325-10]). Negative control siRNA was also used (ON_TARGETplus Non-targeting siRNA#1 [#D-001810-01-05]). 5000 HCC78 cells were plated in 100µL RPM media with 10% FBS serum in a 24-well plate. Invitrogen™’s Stealth™ RNAi Lipofectamine® RNAiMAX (#13778150) optimized protocol for reverse transfection in lung carcinoma cells was followed. 48-hours post-incubation, total RNA was collected using Qiagen’s RNeasy Micro Kit (#74004) according to manufacturer’s protocol.

siRNA Knockdown Assay RNA Transcription Analysis

Reverse transcription quantitative polymerase chain reaction (RT-qPCR) was used to examine changes in RNA expression. RNA was first transcribed into cDNA, pre-amplified to shift qPCR Ct values to an appropriate range, then qPCRRed. Previously designed and verified probes targeting the ROS1 N terminus, ROS1 C terminus (kinase domain), SLC34A2-ROS1 long form, SLC34A2-ROS1 short form, U2AF1 3’UTR, and GAPDH were used in SYBR® Green PCR Master Mix. 2-ΔΔCt analysis was conducted compared to GAPDH and non-targeting siRNA transfection (no siRNA).

TALEN Design and Synthesis

TALENs were designed using the guidelines developed by Eric Kildebeck from the Porteus Lab. TALENs were designed with Cornell University’s TAL Effector
Nucleotide Targeter 2.0 (https://tale-nt.cac.cornell.edu/node/add/talen). The region around U2AF1’s first exon was inputted, the total target site length ranged from 13 to 17 base pairs (bp), the spacer (distance between the binding sites of each individual TALEN) ranged from 14 to 21 bp, and the RVD NN was used in place of NH. Three pairs were chosen to be tested. These pairs are shown in Figure 10. Addgene’s Golden Gate TALEN and TAL Effector Kit 2.0 was used to construct the six TALEN plasmids using Version 2 of their Golden Gate TALEN Assembly protocol.22

![Figure 10](image)

**Figure 10.** The three designed and tested TALEN pairs are shown here. Their target sequences are listed. R1 and L1 comprise one pair, R2 and L2 comprise a second pair, and R3 and L3 comprise another. Each pair meets the design specifications listed in the Methods section.

Donor DNA Design and Synthesis

The donor DNA plasmid provides a template for DNA repair after TALENs create double-stranded breaks in the DNA. Ideally the point mutation would be repaired by changing one nucleotide. However, since successful transfection and repair needed to be detected in a cost-effective and quick fashion, other components were added to the genomic DNA. The designed portion of the donor plasmid is shown in Figure 11. The 5’ arm and 3’ arm are complementary to the area where the gene should be cut, thus setting up for insertion of the sequence into the genome (Figure 12). The actual repair of the mutation is in the complimentary DNA (cDNA) portion. The cDNA is the U2AF1 coding sequence, derived from mature RNA. Wildtype and mutant cDNA-containing
plasmids were created. The nucleotide sequence of the first exon was “wobbled,” meaning it was changed while still preserving the codon sequence in order to prevent homology to the 3’ arm, which could lead to incorrect repair. Furthermore, FLAG and 6xHis tags were added before the stop codon for future protein experiments. Following the stop codon is the polyA sequence necessary for successful completion of RNA transcription. The remaining portion of the donor plasmid was added so successful repair could be identified via fluorescence
microscopy and flow cytometry. The sequence for green fluorescent protein was added, surrounded by loxP sites, so it could be removed in the future with Cre recombinase if desired. Throughout the sequences are restriction enzyme sites for cloning constructability into a bacterial vector.

The vector was constructed with standard cloning techniques. Primers were designed so restriction enzymes could be used to ligate in the various parts of the plasmid previously existing in independent plasmids. Primers were also used to “wobble” the first exon’s sequence.

**Transfection**

Using Invitrogen™’s Neon™ Transfection System (#MPK10096) with R buffer, 1350 V voltage, 10ms width, 3 pulses (based on the NCI-H23 Neon™ Transfection Protocol), HCC78 cells were transfected. 1 million cells in each well of a 12-well plate were transfected with 0.5µg of TALEN Left plasmid, 0.5µg of TALEN Right plasmid, and 2µg of donor DNA plasmid. Cells were grown for two weeks then sorted into 6-well plates and allowed to grow for two more weeks. Cells were single cell sorted into 96-well plates and allowed to grow for another month (see Figure 13). PCR assays are being conducted to verify the faithful insertion of the donor DNA.

**Drug Assays**

5000 HCC78 cells were plated into wells of a 96-well plate. Each well held 300µL total volume including media and drugs. Either TAE684 or crizotinib
were added to the cells. TAE684 was added in the following concentrations: 0, 0.1, 10, 100, 1000, 3000, 10000, and 30000 nM. Crizotinib was added in the following concentrations: 0, 1, 100, 1000, 3000, 100000, and 222222 nM. Drugs were resuspended in DMSO, so 0 nM drug is just DMSO. Cells were incubated for 48 hours; then cell viability was assayed using Life Technologies™'s AlamarBlue® Cell Viability Reagent according to standard protocol. After a 2-hour incubation at 37°C, the assay was evaluated via fluorescence by a plate reader: an excitation wavelength of 570nm and an emission wavelength of 585nm were used.

RNA-seq

Total RNA will be collected using Qiagen’s RNeasy Micro Kit (#74004) according to manufacturer’s protocol. RNA-seq preparation will be done with the NuGEN™’s Ovation® RNA-Seq System V2 (#7102), and samples will be sent to the Stanford Genome Center for sequencing using the Illumina HiSeq 2000.

As for analysis, the following commands will be run on a server with 24 logical cores, 192 GB RAM, 196 GB swap, 28 TiB storage that has Cufflinks v2.0.2 installed:

**STAR**

- Make p190_1 directory to store project
- Move STAR_2.3.0e.Linux_x86_64.tgz to p190_1
- Unzip using `gunzip STAR_2.3.0e.Linux_x86_64.tgz`
- Untar and extract using `tar -xf STAR_2.3.0e.Linux_x86_64.tar`
- Select appropriate genome from [http://cufflinks.cbcb.umd.edu/igenomes.html](http://cufflinks.cbcb.umd.edu/igenomes.html)
- Download genome mm9 using `wget ftp://igenome:
  G3nom3s4u@ussd-ftp.illumina.com/Mus_musculus/
  UCSC/mm9/Mus_musculus_UCSC_mm9.tar.gz` and `tar --xf it`
- Run the genome generating command
  ```bash
  ../../../STAR_2.3.0e.Linux_x86_64/STAR --runMode
genomeGenerate --genomeDir ./ --genomeFastaFiles
../Mus_musculus/UCSC/mm9/Sequence/WholeGenomeFasta/g
  enome.fa --runThreadN 15
  ```
- Make a new test directory and from this directory, and map the reads for each pair of files
  ```bash
  ../../../STAR_2.3.0e.Linux_x86_64/STAR --genomeDir
  ../../../genome/starIndexedGenome/ --runThreadN 15
  ```

### CUFFLINKS

- Change SAM to BAM
  ```bash
  samtools view -bSh Aligned.out.sam -o Aligned.out
  ```
  - `b` means bam output
  - `S` means sam input
  - `h` means include header
  - `o` means output file name
- Sort bam
  ```bash
  samtools sort Aligned.out Aligned.out.sorted
  ```
- Make transcriptome
o in test1 folder,

- `cufflinks -g ./genome/Mus_musculus/UCSC/mm9/Annotation/Genes/genes.gtf -p 2 -o cufflinksOutwGTF ./Aligned.out.sorted.bam`
  - p is number of threads
  - o is output directory

- `cuffmerge -o cuffmergeOutwGTF -g ./genome/Mus_musculus/UCSC/mm9/Annotation/Genes/genes.gtf -s ./genome/Mus_musculus/UCSC/mm9/Sequence/WholeGenomeFasta/genome.fa -p 14 assemblieswGTF.txt`

- `cuffdiff -o cuffdiffOutwGTF -b ./genome/Mus_musculus/UCSC/mm9/Sequence/WholeGenomeFasta/genome.fa -p 14 -L p190+Pax5-,p190,positiveCtrl,negativeCtrl -u ./cuffmergeOutwGTF/merged.gtf ./test2/Aligned.out.sorted.bam ./test4/Aligned.out.sorted.bam ./test5/Aligned.out.sorted.bam ./test3/Aligned.out.sorted.bam ./test1/Aligned.out.sorted.bam`

- `awk '$14=="yes"{print $0}' gene_exp.diff > gene_exp_sig.txt` gives you “significant”ly differentially expressed genes

IV. Results

siRNA Knockdown of U2AF1 and Resulting RNA Expression

In 24 hours, 95% knockdown efficiency of the U2AF1 gene was achieved, and at 48 hours, 90% knockdown remained (data not shown). As shown in Figure 14, with the non-targeting siRNA, the SLC34A2-ROS1 short and long isoforms are
equally expressed. After adding the U2AF1-targeting siRNA, however, the short form is expressed 3-fold higher compared to the long form. There is no significant change in relative ROS1 C (kinase domain) and N termini expression with the knockdown of U2AF1.

**TALENs and Donor DNA Transfection**

Originally lipofection was attempted to insert TALENs and the donor DNA into the cells with Lipofectamine® LTX and Plus Reagent, but it proved lethal to the cells. Various Lonza Nucleofector® Kits were then tried, but without success. Invitrogen™’s Neon™ Transfection System enabled successful TALENs and donor DNA transfection. The cells were genomically edited as evidenced by flow cytometry measurements of GFP fluorescence, which could have only come from the donor DNA. *Figure 15* shows flow cytometry data three weeks post-transfection. Cells were monitored and have been grown for two months to create
Figure 15. Representative flow cytometry data of transfected cells is shown. Plot 1 (left-side) is side-scatter vs. forward-scatter. From here the live cells have been gated and are shown on Plot 2 (right-side), which is FL1 vs. FL2 detection spectra. Cells appearing in the upper left gated area of Plot 2 are GFP positive. (a) Positive control: transfection with Lonza’s pmaxGFP® vector. GFP positive cells are detected. (b) Negative control: transfection with the mutant donor DNA plasmid alone. No GFP positive cells have been detected. (c) Experimental condition: transfection with TALEN L2 DNA, TALEN R2 DNA, and WT donor DNA. A relatively small number of GFP positive cells have been detected.
a stable cell line. Only TALEN pair L2/R2 successfully produced edited cells. Primer assays are currently being performed to confirm the presence of the donor DNA in the appropriate location.

**Drug Assays**

Drug assays were conducted on un-edited HCC78 cells, resulting in the protocol listed in the Methods section. EC50’s for both crizotinib and TAE684 were found within the concentration range being tested, so the same drug concentrations will be used in future experiments. At this time this experiment remains to be done on the stable cell lines, and it will be conducted as soon as the stable cell lines are verified.

**RNA-seq**

RNA-seq analysis has been conducted on mock experimental data, resulting in the protocol listed in the Methods section. At this time this experiment remains to be done on the stable cell lines, and it will be conducted as soon as the stable cell lines are verified.

**V. Discussion**

**siRNA Knockdown Assay**

From the siRNA knockdown assay, it appears that knocking down the U2AF1 gene (both mutated and wild type forms) leads to increased expression of the SLC34A2-ROS1 short form compared to the long form. Looking at the numbers comprising the ratio (not shown), the relative ratio increased because the short form’s expression increased while the long form’s expression stayed the same post-U2AF1 knockdown. This suggests that U2AF1 may affect SLC34A2-ROS1
short form splicing. The hypothesis is that this shift in isoforms will affect cell viability and/or sensitivity to crizotinib, which is a ROS1 kinase inhibitor. This result further motivated the need for endogenously expressed mutant and wildtype U2AF1-expressing cell lines: siRNA knocked down the entire U2AF1 gene, but editing allows comparison between wildtype and heterozygous mutant U2AF1 genotypes specifically.

**TALENs and Donor DNA Transfection**

Successful TALEN editing has been achieved in HCC78. For the HCC78 cell line, it can be said that Life Technologies™ lipofection system and Lonza’s nucleofection kits are not suitable for transfection. Neon™’s transfection system, however, is suitable. Only one of the three pairs of TALENs worked as planned (pair 2), attesting to the need for optimization of TALENs. Based on flow cytometry, a successfully repaired U2AF1 gene should exist in the stable edited-HCC78 cell line. Primer assays are currently being conducted to confirm this with sequencing analysis. Once the stable cell lines (mutant and wildtype U2AF1) have been verified, I can commence the drug assays and RNA-seq experiments.

**Drug Assays and RNA-seq**

After the cell lines have passed the sequence verification process that they are currently undergoing, drug assays and RNA-seq will be conducted on the following experimental conditions: naturally-occurring HCC78, WT-edited HCC78, and mut-edited HCC78. Naturally-occurring HCC78 is a control to see that adding the cDNA does not change HCC78 in any unforeseen way. wt-edited and mut-edited HCC78 conditions will enable a comparison to see if the U2AF1 point mutation has an effect on the dependent variable in question, either drug sensitivity or gene expression. At this time these assays have not been completed.
but are planned to be completed in the following months. RNA-seq should reveal any changes in gene and isoform expression, which is expected since U2AF1 is a component of the spliceosome.

VI. Conclusions and Future Work

The purpose of this thesis was to bioengineer a wt-wt U2AF1 HCC78 cell line and to see the effects of the U2AF1 S34F mutation on HCC78’s gene expression, global splicing events, cell proliferation, and sensitivity to kinase inhibitors. From the siRNA experiment, it can be said that knocking down the U2AF1 gene alters isoform expression of the ROS1 gene, so it is likely that a relationship exists between U2AF1 expression and ROS1-fusion expression. To determine if this relationship is related to the U2AF1 S34F point mutation, a cell line point of comparison has been engineered by genomic editing with TALENs. This cell line should be for all intents and purposes the same HCC78 cell line, except it expresses all wildtype U2AF1 instead of the naturally-occurring mutant phenotype. The next experiments for this cell line are the drug sensitivity experiment and RNA-seq experiment, and methods for those experiments have been established. The hypothesis is that having the U2AF1 mutation will alter the sensitivity of the cells to drugs that target ROS1, such as crizotinib or TAE684. It is expected that the RNA-seq experiments will reveal multiple genes that are differentially expressed compared to their expression with normal U2AF1. In addition to the drug assays and RNA-seq, further work on this project will be continued throughout the year and will include cell proliferation and colony formation assays, xenografts of this cell line into mice to examine tumor formation, and tests of U2AF1’s protein and RNA interactions.
VII. References

11. HCC-78. DSMZ no. ACC 563. Leibniz Institute DSMZ-German Collection of Microorganism and Cell Cultures. https://www.dsmz.de/


